PCT WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro

INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

(51) Internationale Patentklassifikation 6 : G06F 17/30	A1	(11) Internationale Veröffentlichungsnummer: **WO 99/10819** (43) Internationales Veröffentlichungsdatum: 4. März 1999 (04.03.99)

(21) Internationales Aktenzeichen: PCT/DE98/02477

(22) Internationales Anmeldedatum: 24. August 1998 (24.08.98)

(30) Prioritätsdaten:
197 37 145.0 26. August 1997 (26.08.97) DE

(71) Anmelder *(für alle Bestimmungsstaaten ausser US)*: SIEMENS AKTIENGESELLSCHAFT [DE/DE]; Wittelsbacherplatz 2, D–80333 München (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: KOLPATZIK, Bernd [DE/DE]; Unterhachinger Strasse 87, D–81737 München (DE). PFEFFERER, Leo [DE/DE]; Guardinistrasse 46, D–81375 München (DE). SCHAPPERT, Albert [DE/DE]; Flurstrasse 32, D–85244 Röhrmoos (DE).

(74) Gemeinsamer Vertreter: SIEMENS AG; Postfach 22 16 34, D–80506 München (DE).

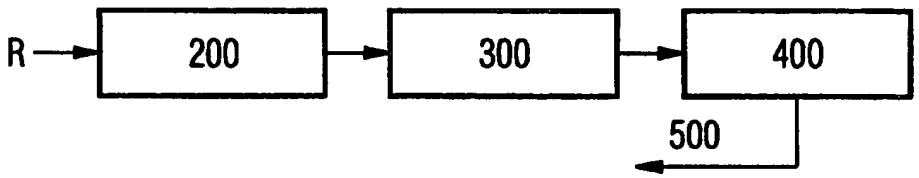
(81) Bestimmungsstaaten: JP, US, europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Veröffentlicht
Mit internationalem Recherchenbericht.
Vor Ablauf der für Änderungen der Ansprüche zugelassenen Frist; Veröffentlichung wird wiederholt falls Änderungen eintreffen.

(54) Title: METHOD AND SYSTEM FOR COMPUTER ASSISTED DETERMINATION OF THE RELEVANCE OF AN ELECTRONIC DOCUMENT FOR A PREDETERMINED SEARCH PROFILE

(54) Bezeichnung: VERFAHREN UND SYSTEM ZUR RECHNERGESTÜTZTEN ERMITTLUNG EINER RELEVANZ EINES ELEKTRONISCHEN DOKUMENTS FÜR EIN VORGEBBARES SUCHPROFIL

(57) Abstract



The invention relates to a method and system for representing the relevance of electronic documents in relation to user–specific search and interest profiles. The relevance of each respective document in relation to specific search profiles is essentially determined by counting words. Documents and search profiles are interpreted as vectors, individual words are considered as vector components and the frequency of words is seen as values of vector components. The document vectors and search profile vectors are projected on a common plane and the angle formed by the vectors is used to measure the conformity of said document in relation to the respective search profile. The results of analysis are represented in three dimensions enabling the documents to be arranged in such a way that similar documents are located next to each other or documents which are relevant to a search profile are arranged close to said search profile. The system can be especially used in searches in computer networks such as Internet or for databank searches and visualization of library contents, archives or complex data stock of all varieties.

(57) Zusammenfassung

Die Erfindung beschreibt ein Verfahren und ein System zur Darstellung der Relevanz elektronischer Dokumente in Bezug auf benutzerspezifische Such– bzw. Interessenprofile. Die Relevanz der jeweiligen Dokumente in Bezug auf bestimmte Suchprofile wird im wesentlichen durch Zählen von Worten bestimmt. Dokumente und Suchprofile werden dabei als Vektoren aufgefaßt, mit den einzelnen Worten als Vektorkomponenten und der Häufigkeit der Worte als Werten der jeweiligen Vektorkomponenten. Die Dokumentenvektoren und Suchprofilvektoren werden in eine gemeinsame Ebene projiziert und der Winkel zwischen den Vektoren dient als Maß für die Übereinstimmung des Dokuments mit dem jeweiligen Suchprofil. Die Analyseergebnisse werden dreidimensional dargestellt und zwar derartig, daß Dokumente so angeordnet werden, daß ähnliche Dokumente beieinander liegen, bzw. Dokumente, die relevant auf ein Suchprofil sind, in der Nähe dieses Suchprofiles angeordnet werden. Angewendet werden kann dieses System insbesondere bei Suchen in Rechnernetzwerken, wie dem Internet bzw. Datenbankrecherchen und zur Veranschaulichung von Bibliotheksinhalten, Archiven oder komplexen Datenbeständen aller Art.

LEDIGLICH ZUR INFORMATION

Codes zur Identifizierung von PCT-Vertragsstaaten auf den Kopfbögen der Schriften, die internationale Anmeldungen gemäss dem PCT veröffentlichen.

AL	Albanien	ES	Spanien	LS	Lesotho	SI	Slowenien
AM	Armenien	FI	Finnland	LT	Litauen	SK	Slowakei
AT	Österreich	FR	Frankreich	LU	Luxemburg	SN	Senegal
AU	Australien	GA	Gabun	LV	Lettland	SZ	Swasiland
AZ	Aserbaidschan	GB	Vereinigtes Königreich	MC	Monaco	TD	Tschad
BA	Bosnien-Herzegowina	GE	Georgien	MD	Republik Moldau	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagaskar	TJ	Tadschikistan
BE	Belgien	GN	Guinea	MK	Die ehemalige jugoslawische Republik Mazedonien	TM	Turkmenistan
BF	Burkina Faso	GR	Griechenland			TR	Türkei
BG	Bulgarien	HU	Ungarn	ML	Mali	TT	Trinidad und Tobago
BJ	Benin	IE	Irland	MN	Mongolei	UA	Ukraine
BR	Brasilien	IL	Israel	MR	Mauretanien	UG	Uganda
BY	Belarus	IS	Island	MW	Malawi	US	Vereinigte Staaten von Amerika
CA	Kanada	IT	Italien	MX	Mexiko		
CF	Zentralafrikanische Republik	JP	Japan	NE	Niger	UZ	Usbekistan
CG	Kongo	KE	Kenia	NL	Niederlande	VN	Vietnam
CH	Schweiz	KG	Kirgisistan	NO	Norwegen	YU	Jugoslawien
CI	Côte d'Ivoire	KP	Demokratische Volksrepublik Korea	NZ	Neuseeland	ZW	Zimbabwe
CM	Kamerun			PL	Polen		
CN	China	KR	Republik Korea	PT	Portugal		
CU	Kuba	KZ	Kasachstan	RO	Rumänien		
CZ	Tschechische Republik	LC	St. Lucia	RU	Russische Föderation		
DE	Deutschland	LI	Liechtenstein	SD	Sudan		
DK	Dänemark	LK	Sri Lanka	SE	Schweden		
EE	Estland	LR	Liberia	SG	Singapur		

Beschreibung

Verfahren und System zur rechnergestützten Ermittlung einer Relevanz eines elektronischen Dokuments für ein vorgebbares Suchprofil.

Die Erfindung bezieht sich auf ein Verfahren und ein System, womit die Relevanz von Dokumenten, wie sie beispielsweise bei einer Internetsuche gefunden werden, bezüglich vorgegebener Interessenprofile dargestellt werden kann.

Die zunehmende elektronische Datenflut in Wissenschaft, Ingenieurwesen und Wirtschaft erschwert das Auffinden und den Zugriff auf relevante, verläßliche und möglichst vollständige Informationen. Bisherige Lösungsvorschläge für Data Mining und Visualisierung großer Informationsmengen, insbesondere von Volltexten und WEB-Seiten, sind häufig weder anwenderfreundlich noch effizient genug für den praktischen Einsatz.

Bestehende Technologien, wie sie z. B. bei Internet Recherchen angewendet werden, beschränken sich zur Zeit noch überwiegend auf die Ausgabe von Texten oder unübersichtlichen Listen von Quellenangaben. Ansätze zur Visualisierung sind zwar in der Literatur dokumentiert, beschränken sich aber entweder auf die Visualisierung wissenschaftlicher Daten, oder vernachlässigen die Aspekte der Erschließung von Informationsbeständen und die Ankopplung an die Visualisierung. Aus dem Artikel von T. Führing, K. Jacoby, R. Michelis, J. Panyr "Kontextgestaltgebung: Eine Metapher zur Visualisierung und Interaktion mit komplexen Wissensbeständen", erschienen in den Proceedings des 4. Internationalen Symposiums für Informationswissenschaft (ISI ´94) Band 16, ist es bekannt eine approximative Einbettung formaler Kontexte in 3D-Informationsräume durchzuführen, deren formale Semantik über den Abstandsbegriff auf der Grundlage des Prinzips "kontextuelle Nähe ≈ räumliche Nähe" definiert wird. Hierdurch ist es möglich binäre formale Kontexte darzustellen.

Aus [1] und [2] ist bekannt, Dokumente hinsichtlich der Relevanz dieser Dokumente bezüglich vorgegebener Schlüsselworte zu analysieren.

Ferner ist aus [3] bekannt, Dokumente hinsichtlich der Häufigkeit des Auftretens eines Schlüsselwortes zu untersuchen.

Der Erfindung liegt die Aufgabe zu Grunde ein Verfahren und ein System für die Veranschaulichung mehrwertiger formaler Kontexte anzugeben.

Diese Aufgabe wird für das Verfahren gemäß den Merkmalen des Patentanspruches 1 und für das System gemäß den Merkmalen des Patentanspruches 13 gelöst.

Bei dem Verfahren zur rechnergestützten Ermittlung einer Relevanz eines elektronischen Dokuments für ein vorgebbares Suchprofil werden mindestens folgende Schritte durchgeführt:

a) es wird das Suchprofil, das mindestens ein Wort umfaßt, erstellt;

b) für jedes Wort des Suchprofils wird die Auftrittshäufigkeit des Wortes in dem elektronischen Dokument bestimmt;

c) unter Verwendung der Auftrittshäufigkeit jedes Wortes wird für das elektronische Dokument ein Ergebnisprofil bestimmt;

d) unter Verwendung des Suchprofils und des Ergebnisprofils des elektronischen Dokuments wird ein Vektor für das Suchprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und ein vorgebbarer Wert ein Wert der Vektorkomponente ist, und ein Vektor für das Ergebnisprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und die entsprechende Häufigkeit ein Wert der Vektorkomponente ist;

e) es wird ein Winkel zwischen dem Vektor des Suchprofils und dem Vektor des Ergebnisprofils bestimmt;

f) unter Verwendung des Winkels wird die Relevanz bestimmt.

Diese Relevanzbestimmung läßt sich mit relativ geringem Rechenaufwand durchführen, so daß viele Suchprofile in bezug auf viele Dokumente analysiert werden können und gleichzeitig ein akzeptables Zeitverhalten erreicht wird.

Das System zur rechnergestützten Ermittlung einer Relevanz eines elektronischen Dokuments für ein vorgebbares Suchprofil weist mindestens folgende Merkmale auf:

a) es ist ein Rechner (COMP) vorhanden, der derart eingerichtet ist, daß folgende Schritte durchführbar sind:
   - es wird das Suchprofil, das mindestens ein Wort umfaßt, erstellt;
   - für jedes Wort des Suchprofils wird die Auftrittshäufigkeit des Wortes in dem elektronischen Dokument bestimmt;
   - unter Verwendung der Auftrittshäufigkeit jedes Wortes wird für das elektronische Dokument ein Ergebnisprofil bestimmt;
   - unter Verwendung des Suchprofils und des Ergebnisprofils des elektronischen Dokuments wird ein Vektor für das Suchprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und ein vorgebbarer Wert ein Wert der Vektorkomponente ist, und ein Vektor für das Ergebnisprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und die entsprechende Häufigkeit ein Wert der Vektorkomponente ist;
   - es wird ein Winkel zwischen dem Vektor des Suchprofils und dem Vektor des Ergebnisprofils bestimmt;
   - unter Verwendung des Winkels wird die Relevanz bestimmt;

b) es ist eine grafische Rechneranzeigevorrichtung (DIS) vorhanden;

c) es sind Mittel zum Zugriff (Z) auf elektronische Dokumente (D) vorhanden.

Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Vorzugsweise werden ein erstes, den Vektor eines Suchprofils repräsentierendes, Element und ein zweites, den Vektor eines Ergebnisprofil eines elektronischen Dokuments repräsentierendes, Element dargestellt.

In einer weiteren Ausgestaltung der Erfindung werden mehrere zweite Elemente, die jeweils einen Vektor eines Ergebnisprofils eines elektronischen Dokuments repräsentieren, derart dargestellt, daß zweite Elemente von elektronischen Dokumenten, welche Dokumente eine Relevanz aufweisen, die kleiner ist als ein Schwellenwert, örtlich näher beieinander dargestellt werden als zweite Elemente von elektronischen Dokumenten, welche elektronische Dokumente eine Relevanz aufweisen, die nicht kleiner ist als der Schwellenwert.

Vorteilhaft wird die Erfindung durch Anwendung einer Winkelfunktion auf die gefundenen Winkel zwischen den Suchvektoren und den Ergebnisvektoren weitergebildet und in Form einer Relevanzmatrix weiterverarbeitet, da diese als Ähnlichkeitsmatrix interpretiert oder auf einfache Weise in eine solche umgewandelt werden kann.

Vorteilhaft wird die Erfindung unter Verwendung einer Ähnlichkeitsmatix weitergebildet, welche aus der Relevanzmatrix abgeleitet wird, und die Ähnlichkeit einzelner Dokumente untereinander angibt. Auf diese Weise läßt sich die Metapher "räumliche Nähe = inhaltliche Nähe" in der graphischen Darstellung sehr einfach realisieren und somit ist bei der Aufbereitung für die Graphik ein geringerer Rechenaufwand erforderlich.

Vorteilhaft wird die Erfindung durch die Anwendung der Kosinusfunktion auf die gefundenen Winkel zwischen den Vektoren

weitergebildet, da der Kosinus von 0° = 1 ist. Somit wird bei einem Übereinanderliegen der Vektoren eine Identität der Dokumente angegeben, was dem Sachverhalt, der durch die Vektoren dargestellt wird, entspricht.

Vorteilhaft wird das erfindungsgemäße Verfahren durch die Anwendung in einem Rechnernetzwerk weitergebildet, da häufig aus Rechnernetzwerken elektronische Dokumente als Suchergebnisse erhalten werden, welche innerhalb eines akzeptablen Zeitabschnitts nicht von Menschen analysiert werden können.

Vorteilhaft wird in einer Weiterbildung der Erfindung als Rechnernetzwerk das Internet verwendet, da das Internet bzw. World Wide Web ein weit verbreitetes Netzwerk darstellt und somit eine hohe Nutzerbasis für das erfindungsgemäße Verfahren vorliegt.

Vorteilhaft wird die Erfindung durch die Verwendung von elektronischen Dokumenten aus Datenbanken weitergebildet, da hierdurch Bibliotheken und andere Datenbanken für elektronische Dokumente sinnvoll, transparent und schnell veranschaulicht werden können.

Vorteilhaft ist ein System bestehend aus einem Rechner einem Display und Mittel zum Zugriff auf elektronische Dokumente, welches das erfindungsgemäße Verfahren und vorzugsweise seine Weiterbildungen ausführt, da die Hardware-Mittel weit verbreitet sind und eine gute Verfügbarkeit dieser Mittel gewährleistet ist. Ebenfalls ist der Zugriff auf elektronische Dokumente durch weitverbreitete Netzzugangsmittel und öffentliche und private Netze gewährleistet.

Im Folgenden werden Ausführungsbeispiele der Erfindung anhand von Figuren weiter erläutert.

Figur 1 zeigt ein Beispiel zur Bildung einer Relevanzmatrix

Figur 2 veranschaulicht weitere Verarbeitungsschritte des Verfahrens

Figur 3 veranschaulicht die Winkelberechnung

Figur 4 zeigt einen Bildschirmausschnitt nach Durchführung des Verfahrens.

Wie Figur 1 zeigt werden bei einer Ausgestaltung des erfindungsgemäßen Verfahrens elektronische Dokumente D1, D2 und DN verwendet und anhand von Suchprofilen P1, P2 und PM, welche fallweise gewichtete Suchbegriffe enthalten, wird die Information, welche in den Dokumenten D1 bis DN enthalten ist, erschlossen. Bei den verwendeten Dokumenten D1 bis DN kann es sich beispielsweise um Dokumente handeln, welche im World Wide Web bei einer Net-Suche gefunden wurden. Bei den Profilen kann es sich um handerstellte bzw. vom Benutzer definierte Suchprofile handeln, welche fallweise an den einzelnen Begriffen Gewichtungen gemäß ihrer Wichtigkeit aufweisen. Ebenfalls ist es denkbar als Profile auch Dokumente zu verwenden. Beispielsweise ist es auch denkbar Suchprofile anhand von Wortstatistiken zu erstellen, welche anhand von Dokumenten durchgeführt werden, die der Bediener für höchst interessant hält und dem Rechner zur Verfügung stellt. Ebenso ist es denkbar Suchprofile unterstützt durch einen fachspezifischen Thesaurus einzugeben. Auch können durch Beobachten des Benutzerverhaltens und durch Lernkomponenten Suchprofile automatisch generiert werden

In einem Bearbeitungsschritt 100 wird die Relevanz zwischen den einzelnen Profilen P1 bis PM und den einzelnen Dokumenten D1 bis DN bestimmt. Vorzugsweise geschieht dies für alle Dokumente und alle Profile, so daß eine Relevanzmatrix R entsteht. Zur Bestimmung der Relevanz wird vorzugsweise die Worthäufigkeit in den Dokumenten ermittelt und übereinstimmende Worte mit den jeweiligen Suchprofilen werden gesucht. Anschließend werden die Suchprofile und die je Dokument und Suchprofil ermittelten Ergebnisprofile als Vektor dargestellt und in der Vektorebene, die durch die Begriffe des Suchvek-

tors aufgespannt wird, wird der Winkel zwischen den Suchvektor und dem Ergebnisvektor bestimmt und als Maß für die Relevanz des Dokumentes das untersucht wurde, verwendet. In Figur 1 ist die Relevanzmatrix R mit Zahlen und Buchstaben versehen, um anzudeuten, wie eine Relevanzmatrix aussehen kann. Waagerecht sind beispielsweise die Profile P1 bis PN aufgetragen und senkrecht die Dokumente D1 bis DN. An den Schnittpunkten der jeweiligen Spalten und Zeilen stehen die Relevanzwerte. Hierdurch wird erstmals ein mehrwertiger formaler Kontext realisiert, wodurch die i-te-Zeile der Matrix R den Relevanzen des i-ten-Dokuments bezüglich aller Profile k entspricht.

Wie Figur 2 weiter zeigt kann die Relevanzmatrix R in Prozeßschritten 200, 300 und 400 weiterverarbeitet werden. Beispielsweise steht über eine Schnittstelle 500 der Zugriff auf Dokumente und Suchprofile und Browser zur Verfügung. In einem ersten Schritt 200 wird beispielsweise aus der Relevanzmatrix eine Ähnlichkeitsmatrix berechnet, wozu aus den Relevanzwerten für einzelne Dokumente mit anderen Dokumenten eine Korrelationsanalyse durchgeführt wird. Bevorzugt wied die Korrelationsmatrix C folgende Rechenschritte durch Berechnung der Korrelationskoeffizienten C_{ik} zwischen den Dokumenten bezüglich der Suchprofile aus der Matrix R durch folgende Schritte bestimmt:

-Normierung der Zeilenvektoren r_i der Matrix R:

$$q_i=(r_i-m_i)$$

mit Mittelwert $m_i=1/N \sum r_i$

Länge q_i und Standardabweichung $\sigma_i=\mathrm{sqrt}(\Sigma(r_i-m_i)^2)$

-Berechnung der Korrelationskoeffizienten zu

$$c_{ik}=\frac{q_i q_k^T}{\sigma_i \sigma_k}$$ und der Matrix C.

-C entspricht dabei in der Form der bisherigen Ähnlichkeitsmatrix, bzw. einer Gegenstands-Gegenstandsmatrix.

Beispielsweise kann der Mechanismus zur Berechnung der Ähnlichkeit durch unterschiedliche Maßnahmen verbessert werden.

-In einem ersten Schritt können beispielsweise Stopwörter eliminiert werden, welche im allgemeinen von der Domäne der Abhandlung des speziellen Dokumentes abhängig sind. In vielen Fällen können dieses Konjunktionen, Artikel, Präpositionen sein, die sicher entfernt werden können, ohne daß dabei der Inhalt des Dokumentes verfremdet wird.

-Fallweise kann es auch möglich sein domänenspezifische Worte zu entfernen, um die Signifikanz des gefundenen Maßes zu verbessern.

-Als weitere Maßnahme kann die Metrik des verwendeten Systems auf wichtige Aspekte der Applikationsdomäne fokussiert werden. In diesem Fall können nur einige wenige Konzepte oder Aspekte der beschriebenen Worte aus domänenspezifischen Thesauri verwendet werden, oder Ontologien.

-Als weitere Maßnahme kann die Unterscheidungskraft des Verfahrens verbessert werden, indem eine umgekehrte Dokumentfrequenzkorrektur eingeführt wird. Bei dieser Methode werden Wortgewichte verwendet, wobei Worte, die in vielen Dokumenten auftreten, mit einem logarithmischen Faktor F gewichtet werden. Dieser Faktor bestimmt sich beispielsweise so, daß F = log(Anzahl der Dokumente D, welche das Wort W_j enthalten/durch die Gesamtzahl der Dokumente). Als Folge dieser Maßnahme erhält man ein wortabhängig gewichtetes Ähnlichkeitsmaß.

In einem Verarbeitungsschritt 300 findet beispielsweise die Umsetzung der Ähnlichkeitsmatrix für eine räumliche Darstellung gemäß dem anfangs zitierten Stand der Technik statt. In einem Verarbeitungsschritt 400 wird gemäß dem Stand der Technik der in Schritt 300 zur Verfügung gestellte Datensatz dreidimensional visualisiert.

-Darstellung der Korrelationsmatrix C durch räumliche Abstände nach einem bekannten Verfahren.

-Anwendung der bekannten Optimierungsalgorithmen zur grafischen Aufbereitung.

-Berücksichtigung der Merkmale in der graphischen Darstellung.

-Ein Dokument ist relevant zu einem Profil, wenn wenigstens ein Wort des Profils einmal im Dokument auftritt.
→ Der Gegenstand "Dokument i" hat das Merkmal "Profil k".

-Visualisierung im 3D-Raum

-VRML: Anwählen der Dokumente und Profile zeigt die Dokument- und Profildateien im Fenster eines Internet-Browsers (z. B.: Netscape).

Der Weg über eine Ähnlichkeitsmatrix, welche aus der Relevanzmatrix abgeleitet wird, ist beim erfindungsgemäßen Verfahren jedoch nicht zwingend erforderlich. Es besteht ebenso die Möglichkeit eines direkten Ansatzes, wobei die Relevanzmatrix R direkt in einen dreidimensionalen Raum umgesetzt wird. Hier wird nicht die Metapher der Ähnlichkeit zwischen Dokumenten und der räumlichen Nähe benutzt, sondern vielmehr die Relevanz eines Dokuments im Bezug auf ein bestimmtes Merkmal in eine räumliche Nähe umgesetzt. Mit der Erfindung wird erstmals die Integration von Textanalyse, Visualisierung und Retrieval in einem System realisiert. Insbesondere wird durch die Erfindung eine neue Verbindungskomponente angegeben, welche aus den Ergebnissen der Dokumentanalyse die Ähnlichkeit von Dokumenten berechnet. Diese Komponente beruht auf einem Korrelationsverfahren, mit welchem die Korrelationsmatrix berechnet wird, welche anschließend im dreidimensionalen Raum auf einem Computerdisplay visualisiert wird. Hierdurch wird erstmals die Veranschaulichung mehrwertiger formaler Kontexte ermöglicht.

Figur 3 veranschaulicht die Berechnung eines Relevanzwertes eines Dokuments in bezug auf ein Suchprofil. Wie bereits beschrieben, werden dazu die Texte des Dokuments und des Suchprofils als Vektoren dargestellt. Wegen einer einfachen übersichtlichen Darstellung wurde hier lediglich ein Suchprofil mit zwei Worten T10 und T20 gewählt. Beispielsweise werden in diesem Fall epidemologische Dokumente untersucht. Der Begriff T10 bedeutet beispielsweise influenza und T20 bedeutet outbreak. DV bezeichnet den Dokumentenvektor und PV bezeichnet den Suchprofilvektor. An den jeweiligen Achsen T10 und T20 ist die Häufigkeit der Worte angegeben. Der Winkel α dient als Maß für die Übereinstimmung des Suchprofilvektors PV und des Dokumentenvektors DV. Insbesondere kann hierfür der Kosinus des Winkels gebildet werden, da bei einer Übereinstimmung der beiden Vektoren der Winkel 0 wäre und damit der Kosinus 1, was einer exakten Übereinstimmung entspräche.

Zur Berechnung des Relevanzwertes eines Dokuments bezüglich eines Profiles folgt nun ein Beispiel:

Gegeben sei ein Dokument:

{Influenza report: Large influenza outbreak reaches Paris.}

Zu diesem Dokument wird ein Dokumentenvektor, dessen Dimensionen durch die Begriffe "influenza, large, outbreak, paris, reaches, report" bestimmt sind definiert. Das Dokument wird bezüglich dieser Dimensionen als Dokumentenvektor

$$d=\{2, 1, 1, 1 , 1, 1\}$$

dargestellt. Die Elemente des Vektors d entsprechen den Worthäufigkeiten der auftretenden Begriffe.

Ähnlich wie für Dokumente und Dokumentenvektoren wird ein Suchprofil definiert,

{influenza, outbreak},

und ein Profilvektor PV, dessen Elemente Gewichtungen der Begriffsdimensionen "influenza" und "outbreak" entsprechen,

PV={1, 1}.

Es wird die Projektion des Dokumentenvektors d auf die Ebene des Profilvektors berechnet und es ergibt sich der projizierte Dokumentenvektor, DV={2, 1}. Anschließend wird cos α zwischen DV und PV als Maß für die Relevanz r des Dokuments bezüglich des Profils definiert:

$$r=\cos\alpha = \frac{<DV,PV>}{\|DV\|\,\|PV\|}.$$

<DV,PV> ist das Skalarprodukt der Vektoren DV und PV, ||.|| ist die Länge eines Vektors.

Für die Beispielvektoren DV und PV ergibt sich somit eine Relevanz des Dokuments bezüglich des Profilvektors von

$$r = \frac{(2+1)}{\sqrt{5}\sqrt{2}} = 0{,}95.$$

Der Spezialfall r=1, bzw. $\alpha=0°$ entspricht der bestmöglichen Relevanz des Dokuments bezüglich des Profils. Ein Wert r=0 ergibt sich bei minimaler Relevanz, bzw. Othogonalität zwischen DV und PV.

Es folgt ein Beispiel zur Berechnung der Korrelationskoeffizienten c_{ik} aus der Relevanzmatrix R:

Gegeben seien zwei Zeilenvektoren r_i und r_k der Matrix R, welche die Relevanzen der Dokumente i und k bezogen auf vier

Profile enthält. Die Vektoren der Zeilen i und k enthalten die Elemente,

$r_i=(0.6,\ 0.2,\ 0.4,\ 0.8)$

und

$r_k=(0.0,\ 0.1,\ 0.3,\ 0.4)$.

Daraus ergeben sich die Mittelwerte

$m_i=0.5,\ m_k=0.2.$

Weiter erhält man

$q_i=r_i-m_i=\ (0.1,\ -0.3,\ -0.1,\ 0.3)$

$q_k=(-0.2,\ -0.1,\ 0.1,\ 0.2),$

mit Längen

$\sigma_i=0.4472,\ \sigma_k=0.3162.$

Für den Korrelationskoeffizienten c_{ik} ergibt sich,

$$c_{ik}=\frac{q_i q_k^T}{\sigma_i \sigma_k}=0.4243.$$

Dieser Koeffizient wird als Maß der Ähnlichkeit von Dokumenten i und Dokument k, bezüglich der vier Profile interpretiert. Die Matrix C hat die Form einer Gegenstands-Gegenstands-Ähnlichkeitsmatrix und kann mit bekannten Verfahren visualisiert werden.

Wie Figur 4 zeigt, kann eine Dokumentenauswertung in bezug auf Interessen bzw. Suchprofile auf einem Bildschirm DIS veranschaulicht werden. Auf dem dargestellten Bildschirmausschnitt sind Dokumente als Würfel und Suchprofile als Kugeln dargestellt. Im einzelnen handelt es sich bei den Suchprofi-

len um summer, Complication, Measles, Chicken-Pox dazu gastro-entritis, Diarrhea, winter, Vaccine illness/outbreak, flu, Mumps. Die Dokumente sind im einzelnen nicht bezeichnet. Durch anklicken eines Dokumentes mit dem Cursor CU wird beispielsweise ein Fenster 10 angezeigt, in welchem der Inhalt des jeweiligen Dokumentes dargestellt wird. Wichtig ist hierbei, daß durch die Anordnung der einzelnen Dokumente zwischen den einzelnen Suchprofilen genau angegeben wird, inwieweit die einzelnen Suchprofile in bezug auf dieses Dokument relevant sind. Bei der erfindungsgemäß durchzuführenden Analyse der einzelnen elektronischen Dokumente können für die einzelnen Suchbegriffe in den jeweiligen Suchprofilen Gewichtungsfaktoren vergeben werden, damit diese beispielsweise abgeschwächt gewichtet werden können, was zu einer geringeren Häufigkeit in bezug auf die Übereinstimmung bestimmter Worte mit den jeweiligen Dokumenten führen würde. Anstatt eines zweidimensionalen Computer Displays DIS können auch dreidimensionale Anzeigevorrichtungen, wie Virtual-Reality-Räume, Head Mounted Display, 3D-Display oder holographisch arbeitende Anzeigen Verwendung finden.

In diesem Dokument sind folgende Veröffentlichungen zitiert:

[1]: US 5 649 193

[2]: US 5 576 954

[3]: US 5 642 518

Patentansprüche:

1.Verfahren zur rechnergestützten Ermittlung einer Relevanz eines elektronischen Dokuments für ein vorgebbares Suchprofil das folgende Schritte umfaßt:

a) es wird das Suchprofil, das mindestens ein Wort umfaßt, erstellt;

b) für jedes Wort des Suchprofils wird die Auftrittshäufigkeit des Wortes in dem elektronischen Dokument bestimmt;

c) unter Verwendung der Auftrittshäufigkeit jedes Wortes wird für das elektronische Dokument ein Ergebnisprofil bestimmt;

d) unter Verwendung des Suchprofils und des Ergebnisprofils des elektronischen Dokuments wird ein Vektor für das Suchprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und ein vorgebbarer Wert ein Wert der Vektorkomponente ist, und ein Vektor für das Ergebnisprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und die entsprechende Häufigkeit ein Wert der Vektorkomponente ist;

e) es wird ein Winkel zwischen dem Vektor des Suchprofils und dem Vektor des Ergebnisprofils bestimmt;

f) unter Verwendung des Winkels wird die Relevanz bestimmt.

2. Verfahren nach Anspruch 1, bei dem jeweils die Relevanz für mehrere Suchprofile und/oder mehrere elektronische Dokumente bestimmt wird.

3. Verfahren nach Anspruch 1 oder 2, bei dem ein erstes, den Vektor eines Suchprofils repräsentierendes, Element und ein zweites, den Vektor eines Ergebnisprofil eines elektronischen Dokuments repräsentierendes, Element dargestellt werden.

4. Verfahren nach Anspruch 3, bei dem mehrere zweite Elemente, die jeweils einen Vektor eines Ergebnisprofils eines elektronischen Dokuments

repräsentieren, dargestellt werden, derart, daß zweite Elemente von elektronischen Dokumenten, welche Dokumente eine Relevanz aufweisen, die kleiner ist als ein Schwellenwert, örtlich näher beieinander dargestellt werden als zweite Elemente von elektronischen Dokumenten, welche elektronische Dokumente eine Relevanz aufweisen, die nicht kleiner ist als der Schwellenwert.

5. Verfahren nach Anspruch 2 bis 4, bei dem unter Verwendung der Relevanzen eine Relevanzmatrix (R) bestimmt wird.

6.Verfahren nach Anspruch 5, bei dem aus der Relevanzmatrix (R) eine Ähnlichkeitsmatrix gebildet wird, indem die Relevanzwerte je elektronischem Dokument (D) zu Relevanzvektoren zusammengefaßt und miteinander korreliert werden und bei dem diese Ähnlichkeitsmatrix für die grafische Darstellung auf einem Rechnerdisplay (DIS) verwendet wird, wobei ein Sinnbild eines ersten elektronischen Dokumentes, welches eine höhere Korrelation mit einem zweiten elektronischen Dokument aufweist als ein drittes, räumlich näher am Sinnbild des zweiten elektronischen Dokumentes dargestellt wird, als das Sinnbild des dritten.

7.Verfahren nach einem der Ansprüche 1 bis 6, bei dem als Winkelfunktion der Kosinus verwendet wird.

8.Verfahren nach einem der Ansprüche 1 bis 7, bei dem als elektronische Dokumente (D) Suchergebnisse einer Suche in einem Rechnernetzwerk verwendet werden.

9.Verfahren nach Anspruch 8, bei dem als Rechnernetzwerk das Internet verwendet wird.

10.Verfahren nach einem der Ansprüche 1 bis 7, bei dem als elektronische Dokumente (D) Dokumente aus einer Datenbank verwendet werden.

11.Verfahren nach einem der vorangehenden Ansprüche, bei dem als Suchprofile (P) elektronische Dokumente (D) verwendet werden.

12.Verfahren nach einem der vorangehenden Ansprüche, bei dem ein auf der Angezeigevorrichtung (DIS) angezeigte Sinnbild mittels einer Eingabevorrichtung der Rechners ausgewählt und/oder der Textinhalt des Dokumentes für das das Sinnbild steht zur Anzeige gebracht wird.

13.System zur rechnergestützten Ermittlung einer Relevanz eines elektronischen Dokuments für ein vorgebbares Suchprofil mit folgenden Merkmalen:

a) es ist ein Rechner (COMP) vorhanden, der derart eingerichtet ist, daß folgende Schritte durchführbar sind:
- es wird das Suchprofil, das mindestens ein Wort umfaßt, erstellt;
- für jedes Wort des Suchprofils wird die Auftrittshäufigkeit des Wortes in dem elektronischen Dokument bestimmt;
- unter Verwendung der Auftrittshäufigkeit jedes Wortes wird für das elektronische Dokument ein Ergebnisprofil bestimmt;
- unter Verwendung des Suchprofils und des Ergebnisprofils des elektronischen Dokuments wird ein Vektor für das Suchprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und ein vorgebbarer Wert ein Wert der Vektorkomponente ist, und ein Vektor für das Ergebnisprofil bestimmt, wobei jedes Wort des Suchprofils eine Vektorkomponente und die entsprechende Häufigkeit ein Wert der Vektorkomponente ist;
- es wird ein Winkel zwischen dem Vektor des Suchprofils und dem Vektor des Ergebnisprofils bestimmt;
- unter Verwendung des Winkels wird die Relevanz bestimmt;

b) es ist eine grafische Rechneranzeigevorrichtung (DIS) vorhanden;

c) es sind Mittel zum Zugriff (Z) auf elektronische Dokumente (D) vorhanden.

14. System nach Anspruch 13, bei dem Auswahlmittel vorhanden sind, zur Auswahl eines Sinnbildes auf der Rechneranzeigevorrichtung (DIS).

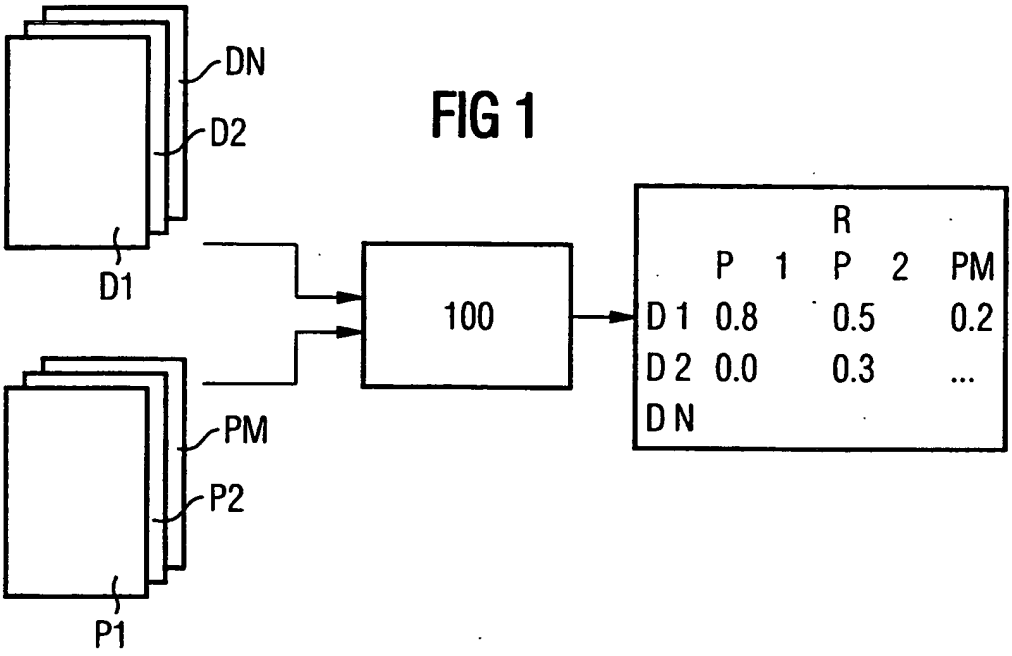


FIG 1

FIG 2

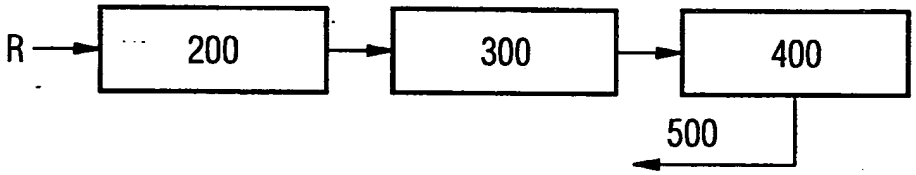


FIG 3

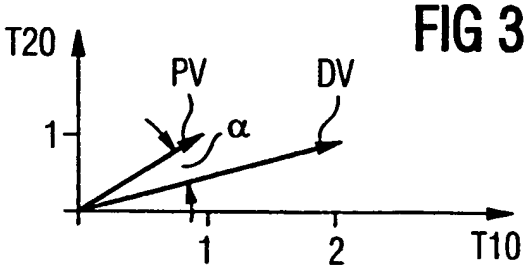
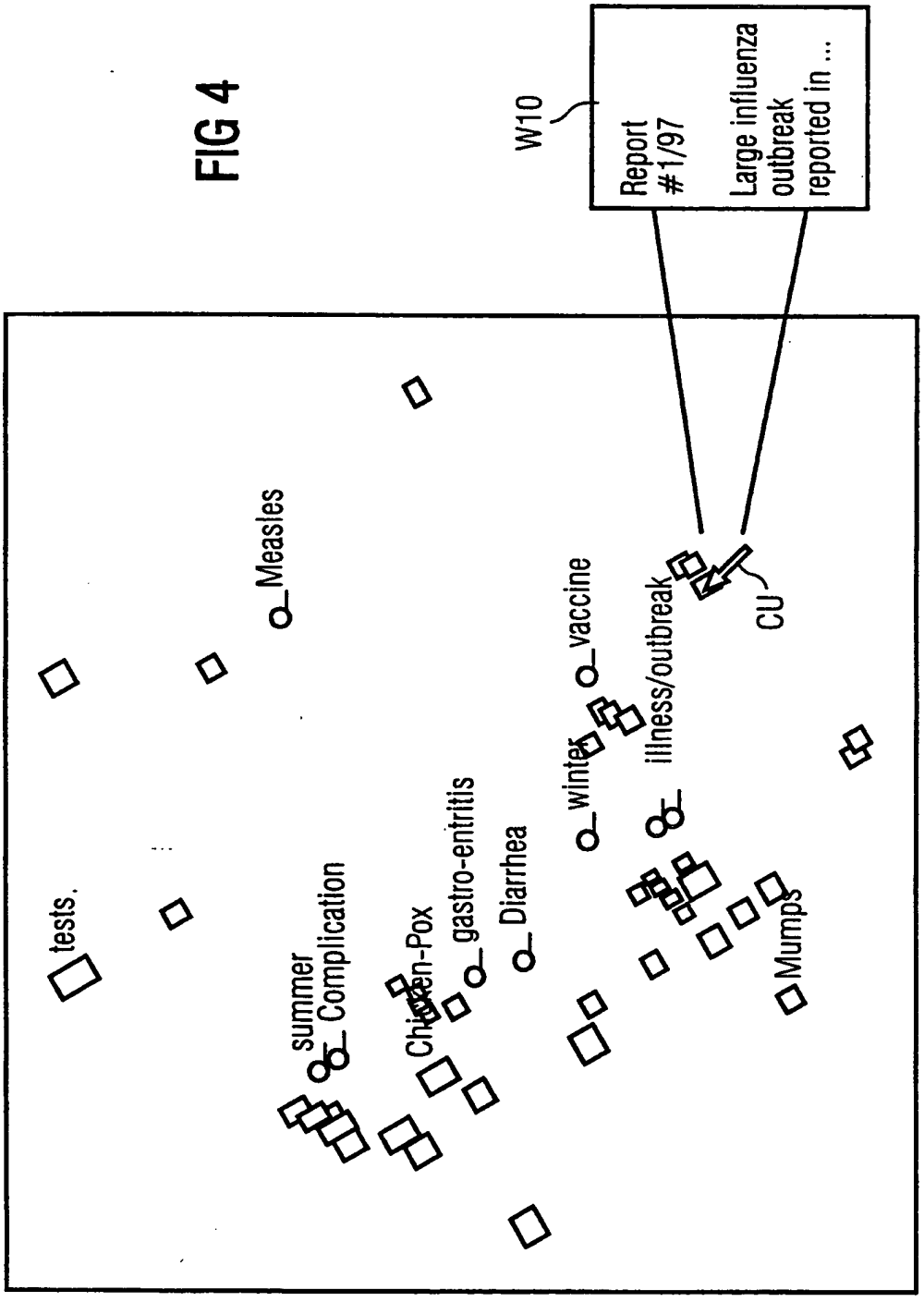


FIG 4


W10
Report #1/97
Large influenza outbreak reported in ...
CU
Measles
vaccine
illness/outbreak
winter
tests.
summer
Complication
Chicken-Pox
gastro-entritis
Diarrhea
Mumps

INTERNATIONAL SEARCH REPORT

Inter ›nal Application No
PCT/DE 98/02477

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SUMNER R G JR ET AL: "An investigation of relevance feedback using adaptive linear and probabilistic models" FIFTH TEXT RETRIEVAL CONFERENCE (TREC-5) (NIST SP 500-238), FIFTH TEXT RETRIEVAL CONFERENCE (TREC-5) (NIST SP 500-238), GAITHERSBURG, MD, USA, 20-22 NOV. 1996, pages 555-570, XP002090102 1997, Gaithersburg, MD, USA, Nat. Inst. Standards & Technol, USA see page 557, line 1 - page 558, line 14 -/--	1-14

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 January 1999	01/02/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Katerbau, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter onal Application No
PCT/DE 98/02477

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	OLSEN K A ET AL: "Visualization of a document collection: the VIBE system" INFORMATION PROCESSING & MANAGEMENT, 1993, UK, vol. 29, no. 1, pages 69-81, XP000574984 ISSN 0306-4573 see page 73, line 6 - page 80, line 13	1-14
X	EGGHE L: "A new method for information retrieval, based on the theory of relative concentration" PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, BRUSSELS, BELGIUM, 5-7 SEPT. 1990, pages 469-494, XP002090103 ISBN 0-89791-408-2, 1989, New York, NY, USA, ACM, USA see the whole document	1-14

1

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/DE 98/02477

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 6 G06F17/30

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 6 G06F

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	SUMNER R G JR ET AL: "An investigation of relevance feedback using adaptive linear and probabilistic models" FIFTH TEXT RETRIEVAL CONFERENCE (TREC-5) (NIST SP 500-238), FIFTH TEXT RETRIEVAL CONFERENCE (TREC-5) (NIST SP 500-238), GAITHERSBURG, MD, USA, 20-22 NOV. 1996, Seiten 555-570, XP002090102 1997, Gaithersburg, MD, USA, Nat. Inst. Standards & Technol, USA siehe Seite 557, Zeile 1 - Seite 558, Zeile 14	1-14

-/--

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[ ] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :

"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist

"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist

"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)

"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht

"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist

"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist

"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden

"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist

"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
15. Januar 1999	01/02/1999

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Katerbau, R

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/DE 98/02477

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	OLSEN K A ET AL: "Visualization of a document collection: the VIBE system" INFORMATION PROCESSING & MANAGEMENT, 1993, UK, Bd. 29, Nr. 1, Seiten 69-81, XP000574984 ISSN 0306-4573 siehe Seite 73, Zeile 6 - Seite 80, Zeile 13	1-14
X	EGGHE L: "A new method for information retrieval, based on the theory of relative concentration" PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, BRUSSELS, BELGIUM, 5-7 SEPT. 1990, Seiten 469-494, XP002090103 ISBN 0-89791-408-2, 1989, New York, NY, USA, ACM, USA siehe das ganze Dokument	1-14